Current Research on HONB-Zipfian Correlations

Leo Behe, Zachary Wheeler
Mentors: Dr. Christie Nelson, Dr. William Pottenger

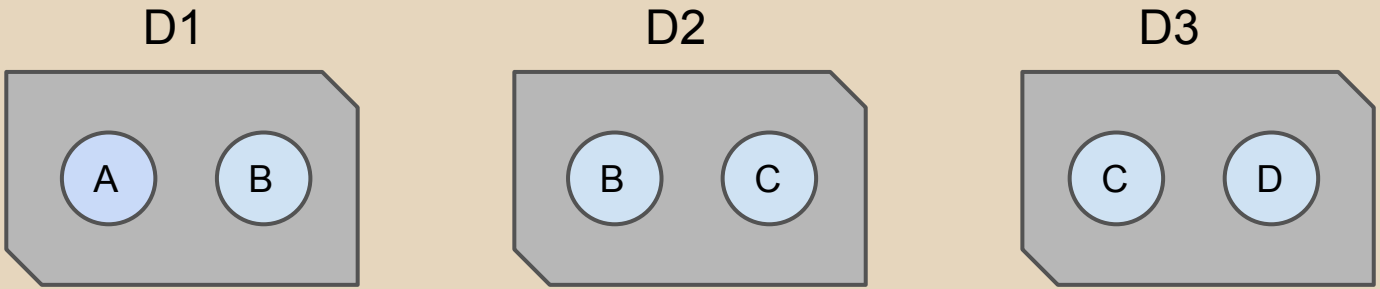
Presented to CCICADA Research Group, 6/18/14

Objectives

- Relational learning often outperforms other methods
  - Why is not well understood
- Prior research suggested a correlation to feature distribution
  - Zipf’s law?
  - Our goal is to investigate

Relational Learning- Textual Example

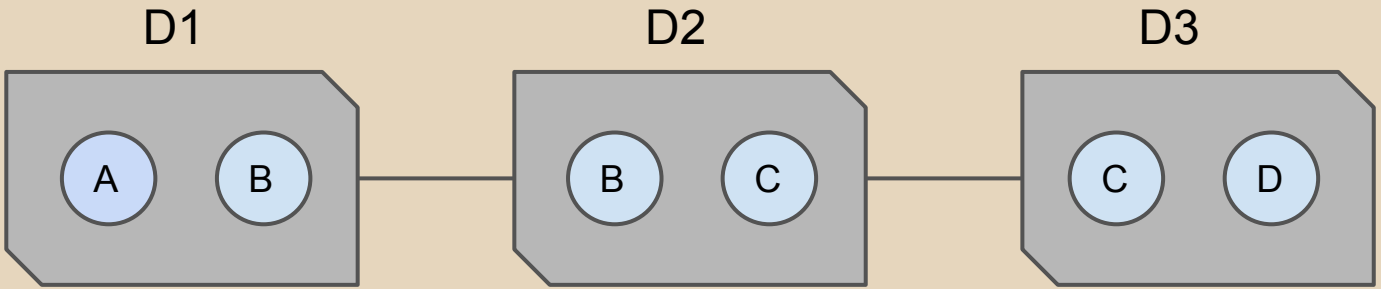
Traditional machine learning:



Note: We will be using documents and words as examples of instances and features in this presentation.

Relational Learning- Example

Relational machine learning:



- Relational learning algorithms have been shown to outperform equivalent IID learning algorithms in certain experiments, especially with small training sets.

Prior Research

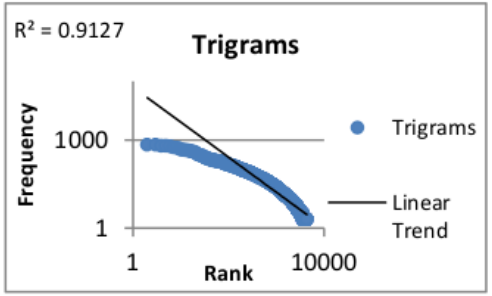


Figure 2: Non-Zipfian Character Trigrams

HONB Accuracy: 44.7%
HONB Precision: 47.0%
HONB Recall: 36.6%

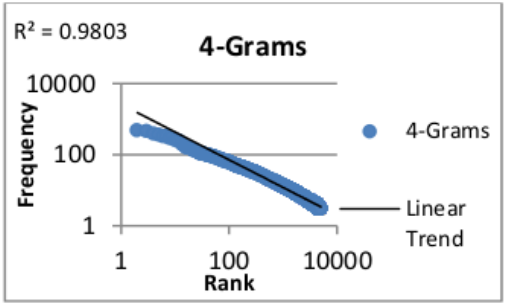


Figure 3: Zipfian Character 4-Grams

HONB Accuracy: 48.8%
HONB Precision: 49.4%
HONB Recall: 42.8%

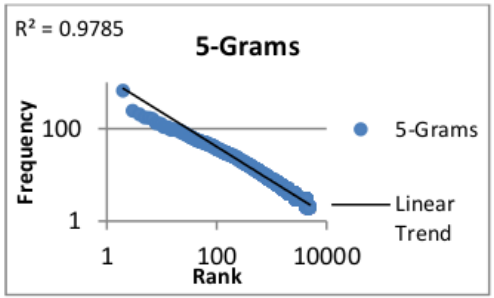


Figure 4: Zipfian Character 5-Grams

HONB Accuracy: 49.5%
HONB Precision: 47.9%
HONB Recall: 46.6%

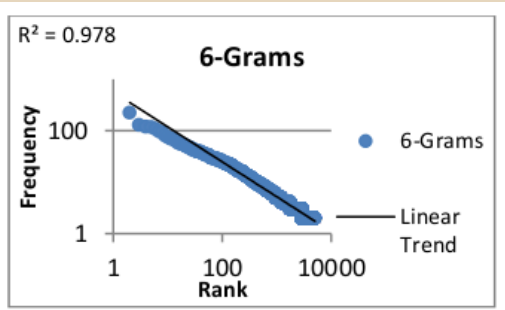


Figure 5: Zipfian Character 6-Grams

HONB Accuracy: 45.1%
HONB Precision: 43.5%
HONB Recall: 43.6%

Datasets

- Ushahidi Haiti earthquake text messages
  - In order to confirm prior results
- Twitter “sentiment” dataset
- California and Washington transportation contracts datasets
  - Contributed by Dr. Williams
- Nuclear detection dataset (numeric)
- World Incidents Tracking System dataset
- White Collar Crime dataset

Tests

- Haiti: n-grams from 2 to 6, feature set of 3000 and 5000 for each n, 20% to 40% with 5% intervals (2528 instances, 5 class, 34% baseline)
- Twitter: n-grams from 2 to 5, feature set of 1000, 3000, and 5000 for each n, 20% to 50% with 10% intervals (1040 instances, 3 classes, 52.9% baseline)
- California: n-grams from 3 to 5, feature set of 3000 and 4000 for each n, 5%, 10%, 20%, 30%, 40% (1221 instances, 3 classes, 44% baseline)

Progress

- Reproduced results of previous experiments
  - Results matched previous results
- Completed tests on Twitter dataset and CA transportation contracts dataset
  - Still analyzing results

Next Steps

- Continue running tests on different types of datasets
- Experiment on non-text-based datasets
- If correlation is supported, formulate hypothetical explanation for it
- Hopefully devise a simple test that can be run on any dataset to determine if relational learning will be effective on it